

IN550 Machine Learning

Rischio atteso, rischio empirico, generalizzazione

Vincenzo Bonifaci

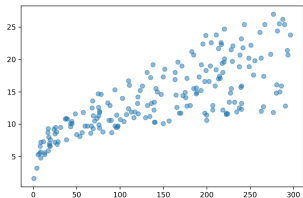
Esempio: Ritorno da investimenti pubblicitari

Input: investimenti pubblicitari via TV, radio e giornali in un mercato (in migliaia di dollari)

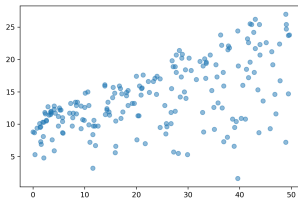
Output: unità di prodotto vendute in quel mercato (in migliaia)

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9
...

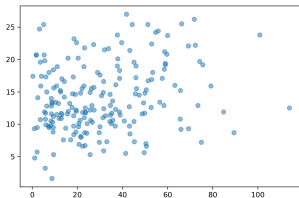
Esempio: Ritorno da investimenti pubblicitari



sales vs. TV



sales vs. radio



sales vs. newspaper

Problemi di predizione: input e output

- Spazio degli input \mathcal{X}
Es.: insieme degli investimenti $\langle \text{tv, radio, giornali} \rangle (\mathbb{R}_+^3)$
- Spazio degli output \mathcal{Y}
Es.: insieme delle possibili quantità di prodotto vendute (\mathbb{R})

Osservati un certo numero di esempi (x, y) , vogliamo trovare una *regola di predizione* (o *ipotesi*)

$$h : \mathcal{X} \rightarrow \mathcal{Y}$$

che ricostruisca in maniera “accurata” la relazione ingresso-uscita

Nei problemi di *regressione* l'output è **quantitativo** (*numerico*)

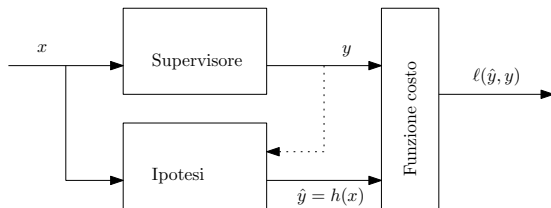
Nei problemi di *classificazione* l'output è **qualitativo** (*categorico*)

Funzione di costo [Loss function]

Come quantificare l'accuratezza di una regola di predizione $h : \mathcal{X} \rightarrow \mathcal{Y}$ su un particolare esempio (x, y) ?

Una *funzione di costo* $\ell : \mathcal{Y} \times \mathcal{Y}_0 \rightarrow \mathbb{R}$ riceve la predizione $\hat{y} = h(x)$ e l'etichetta corretta y , e restituisce un reale nonnegativo

$$\ell(\hat{y}, y) \in \mathbb{R}_+$$



Quadrato dell'errore

$$\ell(\hat{y}, y) \stackrel{\text{def}}{=} (\hat{y} - y)^2$$

Funzione costo 0-1

$$\ell(\hat{y}, y) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } \hat{y} = y \\ 1 & \text{se } \hat{y} \neq y \end{cases}$$

Rischio atteso

Come quantificare l'accuratezza di una regola di predizione $h : \mathcal{X} \rightarrow \mathcal{Y}$ in generale?

Assunzione fondamentale

Gli esempi (x, y) sono generati in modo indipendente da una distribuzione di probabilità (ignota) \mathcal{D} sull'insieme $\mathcal{X} \times \mathcal{Y}_0$

\mathcal{D} è **ignota** poiché è proprio la relazione ingresso-uscita che l'algoritmo cerca di apprendere!

Il *rischio atteso* di una regola di predizione h è

$$\text{RA}(h) \stackrel{\text{def}}{=} \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell(h(x), y)$$

A parole: il rischio atteso di h è il valore atteso della funzione di costo, quando gli esempi sono generati dalla distribuzione \mathcal{D}

Il problema del machine learning supervisionato

Problema del machine learning supervisionato

Fissata una distribuzione (ignota) \mathcal{D} su $\mathcal{X} \times \mathcal{Y}_0$, cerca una regola di predizione che minimizzi il rischio atteso:

$$\underset{h}{\text{minimize}} \text{RA}(h) \left(\equiv \underset{h}{\text{minimize}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell(h(x), y) \right)$$

Il rischio atteso dipende dalla distribuzione ignota \mathcal{D}

Come possiamo minimizzarlo, visto che non conosciamo \mathcal{D} ?!

Non conosciamo \mathcal{D} ma abbiamo degli *esempi* dalla distribuzione \mathcal{D}

Il *rischio empirico* di h sugli esempi $S = \{ (x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)}) \}$ è

$$\text{RE}_S(h) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m \ell(\hat{y}^{(i)}, y^{(i)})$$

dove $\hat{y}^{(i)} = h(x^{(i)})$.

L'idea è di usare il rischio empirico come *surrogato* del rischio atteso

Empirical Risk Minimization (ERM)

Dato un insieme di esempi S (generati da \mathcal{D}), cerca una regola di predizione che minimizzi il rischio empirico su S :

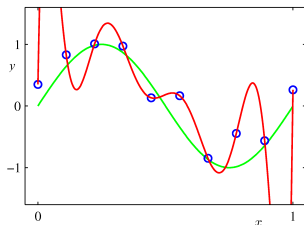
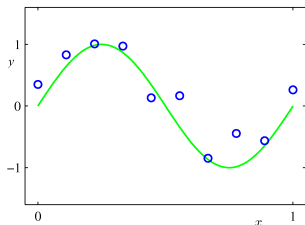
$$\underset{h}{\text{minimize}} \text{RE}_S(h) \left(\equiv \underset{h}{\text{minimize}} \frac{1}{m} \sum_{i=1}^m \ell(\hat{y}^{(i)}, y^{(i)}) \right)$$

L'insieme S di esempi osservati dal learner è detto *training set*

Applicando l'ERM, il problema del learning supervisionato è rimpiazzato da un problema di ottimizzazione nello spazio delle regole

Il sovradattamento (overfitting)

Il principio ERM può fallire gravemente senza le dovute cautele!



In questo esempio, la regola scelta (la funzione rossa) è *sovradattata* ai dati (*overfitting*):

“Spiega” perfettamente le osservazioni, **ma non è un buon modello** della distribuzione da cui i dati sono generati (funzione verde + rumore)

Il suo rischio empirico è nullo, ma il suo rischio atteso è alto

ERM con una classe di ipotesi ristretta

Un approccio al problema dell'overfitting consiste nel **limitare** l'insieme delle possibili regole di predizione (ipotesi)

Anziché considerare la classe $\mathcal{Y}^{\mathcal{X}}$ di **tutte** le funzioni da \mathcal{X} a \mathcal{Y} , consideriamo solo una sua sottoclasse \mathcal{H} (*insieme delle ipotesi*)

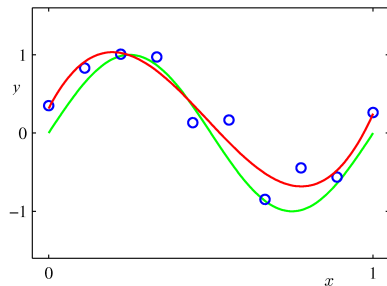
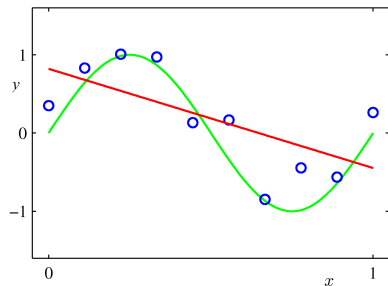
Applichiamo il principio ERM **restringendoci alle ipotesi in \mathcal{H}** :

Empirical Risk Minimization (ERM) – versione ristretta

$$\underset{h \in \mathcal{H}}{\text{minimize}} \text{RE}_S(h)$$

- La classe \mathcal{H} può incorporare la conoscenza pregressa del problema considerato, limitando la *complessità* delle ipotesi
- La classe \mathcal{H} introduce un *pregiudizio (bias) induttivo*: tutte le regole **non** in \mathcal{H} sono scartate a priori

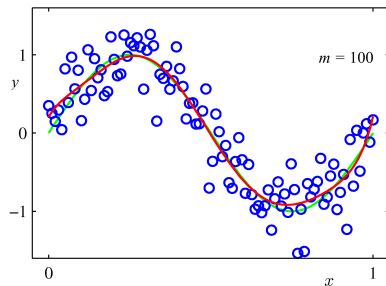
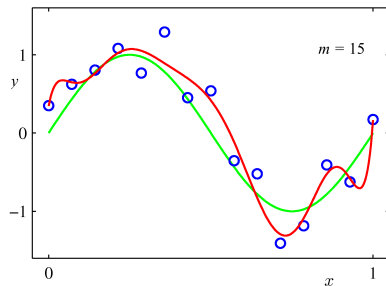
Compromesso bias-varianza



Fitting di un polinomio di grado 1 (sinistra) e di grado 3 (destra)

- Modelli più semplici hanno più bias (possono esibire *underfitting*)

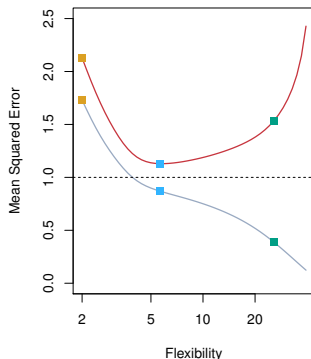
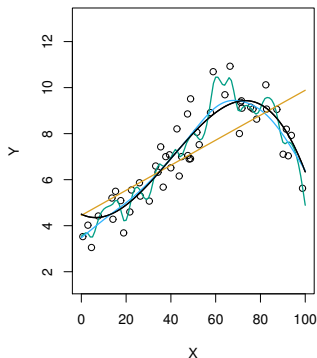
Compromesso bias-varianza



Fitting di un polinomio di grado 9 con 15 esempi (sinistra) e con 100 esempi (destra)

- Modelli più complessi hanno più varianza (richiedono più esempi)

Compromesso bias-varianza



- Sinistra: I dati sono generati sommando la curva nera con un termine di rumore
Le altre curve rappresentano regressioni polinomiali di grado 1, 5, e 23
- Destra: La curva grigia rappresenta il rischio empirico
La curva rossa rappresenta il rischio atteso

Regressione lineare

Nella *regressione lineare*, l'insieme delle ipotesi è l'insieme \mathcal{H}_{lin} delle funzioni **lineari** (affini) da $\mathcal{X} \equiv \mathbb{R}^d$ a $\mathcal{Y} \equiv \mathbb{R}$:

$$h \in \mathcal{H}_{lin} \Leftrightarrow h(x) = w_0 + w_1x_1 + \dots + w_dx_d \quad (w_0, \dots, w_d \in \mathbb{R})$$

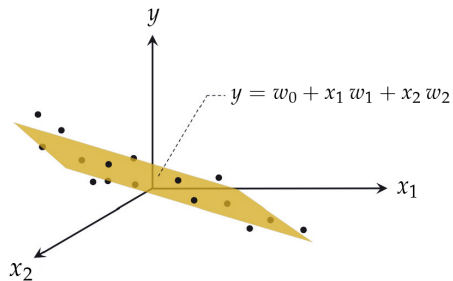
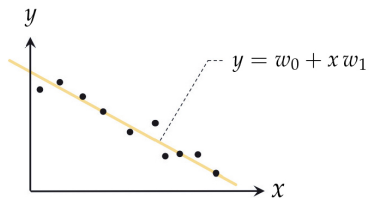
Useremo spesso la convenzione $x_0 \stackrel{\text{def}}{=} 1$, così da poter scrivere $h(x) = w^\top x$

- w_0 è l'*intercetta* (valore previsto dal modello quando x è nullo)
- w_k è il *coefficiente* che esprime la dipendenza di $h(x)$ dalla k -esima componente di x

Una funzione di costo comunemente utilizzata è quella quadratica:

$$\ell(\hat{y}, y) = (\hat{y} - y)^2$$

Regressione lineare

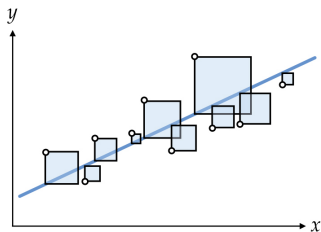
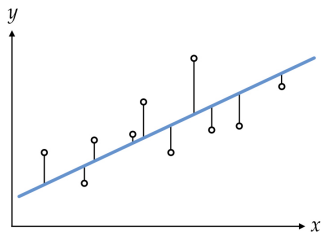


ERM per la regressione lineare

Nella regressione lineare con costo quadratico, il rischio empirico è dato dall'*errore quadratico medio* [*mean squared error*]:

Mean Squared Error (MSE)

$$\text{RE}_S(h) = \frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 = \frac{1}{m} \|Xw - y\|^2$$



ERM per la regressione lineare

Il minimizzatore del rischio empirico qui è esprimibile in forma chiusa:

$$w^* = \left(\sum_{i=1}^m x^{(i)} x^{(i)\top} \right)^{-1} \left(\sum_{i=1}^m y^{(i)} x^{(i)} \right) = (X^\top X)^{-1} X^\top y$$

Infatti (si dimostra) deve soddisfare le cosiddette **equazioni normali**:

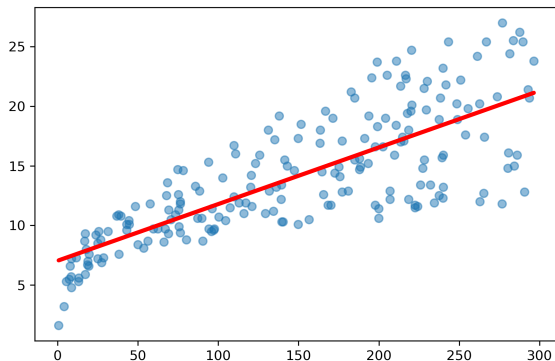
Equazioni normali

Se w^* minimizza l'errore quadratico medio, allora

$$X^\top X w^* = X^\top y$$

Nella pratica, w^* è calcolata con metodi numerici di fattorizzazione (Singular Value Decomposition – SVD), più stabili rispetto alla formula di inversione e che non richiedono l'esistenza dell'inversa

Esempio: regressione di sales su TV



$$\text{sales} \approx w_0 + w_1 \cdot \text{TV}$$

- Intercetta $w_0 = 7.03 \Rightarrow 7030$ unità di prodotto vendute senza investimenti
- Coefficiente $w_1 = 0.047 \Rightarrow 47$ unità di prodotto in più ogni 1000\$ di pubblicità in TV

Come valutare la qualità del modello?

Attenzione

qualità del fit su S (rischio empirico)
 \neq
qualità del modello (rischio atteso)

Possiamo stimare il rischio atteso di un'ipotesi h utilizzando un insieme di esempi di test T (*test set*) **indipendenti** da h

Con sufficienti esempi, il rischio empirico su T sarà una buona stima del rischio atteso:

$$\text{RE}_T(h) \approx \text{RA}(h)$$

Infatti si ha $\mathbb{E}[\text{RE}_T(h)] = \text{RA}(h)$ (esercizio)

Training set e test set

In pratica, si avrà un solo insieme di dati a disposizione

Separiamo a caso i dati di esempio a nostra disposizione in due insiemi S (training) e T (test)



- Il *training set* S è usato per trovare l'ipotesi h col miglior fit:

$$\underset{h \in \mathcal{H}}{\text{minimize}} \text{RE}_S(h)$$

- Il *test set* T è usato per stimare il rischio atteso di h :

$$\text{RE}_T(h) \approx \text{RA}(h)$$

- La separazione è necessaria affinché gli esempi usati per stimare $\text{RA}(h)$ siano indipendenti da h
- La separazione deve essere casuale, affinché S e T seguano la stessa distribuzione
- È un **errore** usare gli esempi di test per fare il training del modello, o usare gli esempi di training per valutare la qualità del modello!

Riepilogo dell'apprendimento ERM

In ogni metodo di apprendimento conforme al principio ERM:

- 1 Si assume una classe di ipotesi \mathcal{H}
- 2 Si assume una funzione di costo ℓ
- 3 Dato un training set di m esempi, attraverso un algoritmo di ottimizzazione si sceglie $h \in \mathcal{H}$ in modo da minimizzare

$$\frac{1}{m} \sum_{i=1}^m \ell(h(x^{(i)}), y^{(i)})$$

- 4 Il rischio atteso di h viene stimato attraverso un test set
- 5 L'ipotesi h viene utilizzata per le predizioni successive:
 - Per ogni nuovo input x' , la predizione è $h(x')$

L'*ipotesi Bayesiana* h^* è quella che fra tutte minimizza il rischio atteso:

$$h^* = \operatorname{argmin}_{h \in \mathcal{Y}^{\mathcal{X}}} \operatorname{RA}(h)$$

Poiché

$$\operatorname{RA}(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell] = \mathbb{E}_x[\mathbb{E}_{y|x}[\ell|x]]$$

abbiamo che per ogni x , $h^*(x)$ minimizza $\mathbb{E}_{y|x}[\ell|x]$
(ciascun $\hat{y} = h^*(x)$ può essere scelto separatamente in base a x)

$\operatorname{RA}(h^*)$ è detto *rischio bayesiano*

Ipotesi Bayesiana: esempio

Nel caso della regressione con costo quadratico, $\ell = (h(x) - y)^2$, quindi:

$$\begin{aligned}\mathbb{E}_{y|x}[(h - y)^2|x] &= \mathbb{E}[h^2 - 2hy + y^2|x] \\ &= h^2 - 2h \mathbb{E}[y|x] + \mathbb{E}[y^2|x] \\ &= h^2 - 2h \mathbb{E}[y|x] + (\mathbb{E}[y|x])^2 + \text{Var}[y|x] \\ &= (h - \mathbb{E}[y|x])^2 + \text{Var}[y|x]\end{aligned}$$

Il secondo termine non dipende da h

Il primo termine è minimizzato se $h(x) = \mathbb{E}[y|x]$

Ipotesi Bayesiana per la regressione con costo quadratico

$$h^*(x) = \mathbb{E}[y|x]$$

Teorema (Decomposizione bias-varianza del rischio atteso)

Se $\ell(h(x), y) = (h(x) - y)^2$, vale la decomposizione

$$\text{RA}(h) = (\text{bias}_h)^2 + \text{varianza}_h + \text{rischio bayesiano}$$

dove:

- $\text{bias}_h \stackrel{\text{def}}{=} \mathbb{E}[h(x) - h^*(x)]$
- $\text{varianza}_h \stackrel{\text{def}}{=} \text{Var}[h(x)] = \mathbb{E}[(h(x) - \mathbb{E}[h(x)])^2]$
- $\text{rischio bayesiano} \stackrel{\text{def}}{=} \text{RA}(h^*) = \mathbb{E}[(h^*(x) - y)^2]$
- $h^*(x) = \mathbb{E}[y|x]$

Decomposizione del rischio atteso in generale

Teorema (Decomposizione stima-approximazione del rischio atteso)

Sia \mathcal{H} una qualunque classe di ipotesi.

Se $h \in \mathcal{H}$, $\bar{h} = \operatorname{argmin}_{h \in \mathcal{H}} \operatorname{RA}(h)$, $h^* = \operatorname{argmin}_{h \in \mathcal{Y}^{\mathcal{X}}} \operatorname{RA}(h)$, allora

$$\operatorname{RA}(h) = \underbrace{\operatorname{RA}(h) - \operatorname{RA}(\bar{h})}_{\text{err. stima}} + \underbrace{\operatorname{RA}(\bar{h}) - \operatorname{RA}(h^*)}_{\text{err. approssimazione}} + \underbrace{\operatorname{RA}(h^*)}_{\text{rischio bayesiano}}$$

dove:

- $\operatorname{RA}(h) - \operatorname{RA}(\bar{h}) \geq 0$ (*errore di stima*)
- $\operatorname{RA}(\bar{h}) - \operatorname{RA}(h^*) \geq 0$ (*errore di approssimazione*)
- $\operatorname{RA}(h^*) \geq 0$ (*rischio bayesiano*)

NB. La decomposizione bias-varianza non è un caso particolare della decomposizione stima-approximazione, anche se qualitativamente simile.